TISMIR

Rao, P., Vinutha, T. P., & Rohit, M. A. (2020). Structural Segmentation of Alap in Dhrupad Vocal Concerts. *Transactions of the International Society for Music Information Retrieval,* 3(1), pp. 137–152. DOI: https://doi.org/10.5334/tismir.64

RESEARCH

# Structural Segmentation of Alap in Dhrupad Vocal Concerts

Preeti Rao, Thallam Prasad Vinutha and Mattur Ananthanarayana Rohit

Dhrupad vocal concerts exhibit a temporal evolution through a sequence of homogeneous sections marked by shared rhythmic characteristics. In this work, we address the segmentation of a concert audio's unmetered improvisatory section into musically meaningful segments at the highest time scale. Motivated by the distinct musical properties of the sections and their corresponding acoustic correlates, we compute a number of features for the segment boundary detection task. Both supervised and unsupervised approaches are tested using a dataset of commercial performance recordings that is manually annotated. The dataset is augmented suitably for training and testing of the models to obtain new insights about the relevance of the different rhythmic, melodic and timbral cues in the automatic boundary detection task. We also explore the use of a convolutional neural network trained on mel-scale magnitude spectrograms for the boundary detection task to observe that while the implicit musical cues are largely learned by the network, it is less robust to deviations from training data characteristics. We conclude that it can be rewarding to investigate knowledge driven features on new genres and tasks, both to achieve reasonable performance outcomes given limited datasets and for drawing a deeper understanding of genre characteristics from the acoustical analyses.

**Keywords:** Structural segmentation; Indian raga music; Dhrupad; Rhythm and timbre features

## 1. Introduction

Musical structure refers to the 'grouping', or the manner in which music is segmented, at a whole variety of levels from groups of a few notes up to the large-scale form of the work (Clarke, 1999). The relationships are created by the temporal order, repetition, homogeneity or contrast of musical aspects. Music structure analysis from audio is an important topic of research in Music Information Retrieval (MIR). However, much of this research has been restricted to Western or popular music cultures and does not generalize easily due to the high dependence of musical structure characteristics on the culture and genre.

In this work, we study the structural segmentation of concerts of the North Indian vocal genre, *Dhrupad*. In particular, we investigate unsupervised and supervised methods for the detection of structural boundaries in the elaborate improvised section of the concert known as the *alap*. Given that the genre has received little attention in MIR, even though considerable musicological scholarship is available, we test existing automatic methods for structural segmentation while exploring new approaches motivated by the characteristics of the music tradition. A top-down system design involving musicology and higher-level culture-specific perspectives can also provide new insights about performance practice over that possible with purely data-driven methods (Serra, 2011).

Audio segmentation is crucial for MIR applications like fast navigation, finding repetitive structure in music or even for the task of music transcription (Klapuri et al., 2001). Metadata supplied with commercial CDs or performance audios on the internet provides information about the musicians and, possibly, about the number and durations of the constituent music pieces. However, information about the section boundaries within a piece is rarely specified. Segmentation can also facilitate other more complicated tasks like section labelling, audio thumbnailing that extracts short representative clips (Bartsch and Wakefield, 2005), and music summarization that stitches these thumbnails to aid rapid browsing (Cooper and Foote, 2003; Peeters, 2003). An Indian art music concert typically involves a solo performer with accompanying instrumentalists. It is generally extempore in nature and lasts for a long duration, even up to a few hours. The overall concert structure stems from distinct sections of specific musical characteristics, organized hierarchically and lasting several minutes each, with the approaching section boundaries cued by the performer in various ways. The identification of section boundaries would therefore be a major step towards rich transcription for pedagogy and musicology research (Widdess, 1994).

The present work addresses the automatic detection of the major section boundaries in Dhrupad alap performance recordings using musicologically motivated

Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, India

Corresponding author: Preeti Rao (prao@ee.iitb.ac.in)

representations of the acoustic cues to section change. Primarily intended to explore the potential of current MIR techniques to a musicologically interesting task in an under-researched genre, this exercise brings out the role of diverse musical cues in modeling section boundaries. We present a brief overview of the music tradition followed by a literature review of music segmentation methods. In Section 2, we present our dataset and annotation procedure. Acoustic features associated with known and observed musical characteristics are discussed in Section 3. Segmentation approaches, experiment conditions and models are explained in Section 4 and the results of boundary detection from the different approaches presented in Section 5. Finally, Section 6 summarizes the work and points to some future tasks.

### *1.1 Music Background*

Music exhibits a rich structure at multiple time scales (McFee et al., 2017). One of the attributes distinguishing an ordinary sound sequence and a musical piece is this intricate structure. In a Hindustani music concert, the artist selects a particular melodic framework (*raga*) appropriate for the mood of the occasion and presents a pre-composed piece called *bandish*, including an elaborate improvisation adhering to the raga rules. Raga rules specify the number of pitch intervals in an octave, their hierarchy and precise intonation as well as the various ascending/descending phrasal contexts (Widdess, 2013). As such, raga delineation is a highly structured activity.

Dhrupad is the ancient North Indian classical style of singing, rendered with the *tambura* as the accompanying drone and the *pakhawaj* as the percussive accompaniment. The word *Dhrupad* refers to the compositional form as well as the genre of Dhrupad which includes singing of *raga alap* (an improvised performance based on the raga grammar) and *pada* (a composition interspersed with *layakari*, its melodic-rhythmic improvisation in performance) (Widdess, 1994). Thus a Dhrupad performance is an elaborate exploration of the raga that starts with an improvised, un-metered alap and is followed by the selected composition with lyrics sung to the accompaniment provided by the *pakhawaj*. A raga performance, that lasts up to an hour, is subdivided unequally into the improvised and composed sections with the former taking up even up to 80 percent of the performance time. The improvisation or alap, is further divided into *alap-proper* or *vilambit alap* (slow and devoid of pulsation), the *jod* or *madhya alap* (medium and steady pulsation) and the *jhala* or *drut alap* (faster pulsation). In some concerts, *jalad jod* or *jalad jhala* (faster versions of jod and jhala respectively) are present as well (Clayton, 2001). Hence, the number of sections in an alap is three or more. **Figure 1** shows the back cover of an audio CD of a Dhrupad performance, where information about the total duration of the alap, and about each of the two composed sections, is provided without further break-down of each of the alap, jod and jhala subsections.

Throughout the alap, the singing comprises phrases made up of syllabes such as *na*, *re*, *te* and *di*, accompanied only by the steady drone. The customary vocal range of Dhrupad is two to two and a half octaves, and the alap begins around the middle octave tonic. The vocalist explores the raga, note by note, sometimes reaching down to the tonic of the lower octave. After exploring the lowest octave, the rendition moves up into the middle octave and ultimately ascends up to the tonic of the highest octave (Wade, 2001). This gradual, progressive melodic ascent is the characteristic of each of the alap-proper and jod sections of the Dhrupad alap. A musical description of the different sections of a Dhrupad alap is presented in **Table 1**. The table also mentions the typical cues that the performer uses to indicate the major structural boundaries, the temporal scale of interest in this work. The hierarchical organization of the concert implies the existence of smaller segment boundaries that are similarly cued. For example, the musician utters a specific melodic motif *noom* at 'the transition of a melodic thought' (Wade, 2001). This is especially significant when the concert involves a pair of singers (an arrangement not uncommon in the Dhrupad style) who alternate with each other as though engaged in a conversation. Further, a specific rhythmic phrase called *mohra* comprising the syllable sequence *ra-na-na-na-na ta-na-tom-ta-na-na* signals a new section boundary when it is rendered at a new higher speed. The section boundaries of interest to us, as indicated in **Table 1** are therefore cued by the *noom* and mohra, periods of silence, and abrupt changes in pulsation (via changing syllable rate). While the sections and their characteristics are distinctive of Dhrupad, it must be noted that the genre 'also comprehends incredible diversity as cultivated by individual musicians' (Wade, 2001).


**Figure 1:** Back cover of a Dhrupad concert audio CD by Pt. Nirmalya Dey.

**Table 1:** Dhrupad alap sections with descriptions.

| Section | Musical characteristics |
|---|---|
| Alap-proper | Rhythm free, slow and elaborate development of raga notes and phrases. A wide melodic range is spanned with focus gradually shifting from middle octave tonic to lower and then the higher octave. The melodic glide *noom* and *mohra* phrase serve as boundary cues. |
| Jod | Introduction of regular and slow pulsations via syllable rate. Melodic development and boundary cues similar to Alap-proper. |
| Jhala | Pulsation accelerates indicating climax. Syllable articulation more regular. The melodic range spanned is relatively narrow. |

### *1.2 Structural Segmentation Literature Review*

Structural boundaries in music are based on changes in the musical attributes of melody, timbre, rhythm or harmony. This motivates the choice of acoustic features used in a segmentation task. Perceptually, timbre represents the sound quality, by which humans distinguish for example different instruments. Chroma features, or pitch class profiles, capture the melodic and harmonic content of a music piece. The relative strengths of each of the 12 notes of the equal-tempered scale characterize both the melody and the harmony in Western music. Timbre and harmony based segmentation methods are more common in Western music and have been used by many researchers to segment popular music into intro-chorus-verse-outro sections (Dannenberg and Goto, 2008; Paulus et al., 2010). Another widely used feature to characterize the global timbre of audio is the set of Mel Frequency Cepstral Coeffients (MFCCs) that parameterize relative sound levels in critical frequency bands (Logan, 2000; Foote and Cooper, 2003). Recently, Allegraud et al. (2019) reviewed the challenges inherent in the segmentation of large-scale structure in the classical sonata form and proposed methods exploiting music theory with computed melodic, harmonic and rhythmic features to characterize the evolving structure as a sequence of recurring states.

Rhythmic cues have been exploited for segmentation of Chinese popular music by Jensen et al. (2005). Unlike in Western music, a beat is not prominently present in some music styles. Hence, the autocorrelation of an accent curve representing the note onsets has been used to get a more robust feature of rhythm over alternate methods based on inter onset interval histograms (IOIH) that rely on a precise location of onsets (Dixon, 2001). A low-dimensional rhythmic representation that captures tempo was used effectively for Western classical music by Grosche et al. (2010). Jensen (2006) incorporated timbre, as well as chroma features along with rhythmic features in the segmentation of various styles of music with visualization using the timbregram and chromagram, confirming that multiple music dimensions are needed to account for the diversity inherent in music. Similarly, annotation principles, segmentation approaches and features were examined for structural segmentation of Chinese traditional Jingju music by Tian and Sandler (2016).

Different approaches are possible for segmenting music into sections – a homogeneity-based approach that locates the sections consistent in some musical aspect, a novelty-based approach that detects a sudden change in musical properties, or a repetition-based method that identifies the recurring patterns in a piece of music. A review by Paulus et al. (2010) of structure analysis from the perspective of segmentation and grouping of similar sections indicates the prominent place of unsupervised approaches. A particularly popular method, introduced by Foote (2000), locates the boundary between contrasting musical features via a self-distance matrix (SDM). Supervised approaches for pop/rock songs utilizing difference features for characterizing changes in musical aspects like timbre, harmony, melody, rhythm have been explored by Turnbull et al. (2007). In this work, they use a boosted decision stump (BDS) classifier that is trained to predict boundary/non- boundary frames. More recently, Ullrich et al. (2014) applied convolutional neural networks (CNN) trained similarly on mel-scaled magnitude spectrograms on the SALAMI structural annotation dataset that spans a large variety of genres (Smith et al., 2011).

Reviewing structural segmentation in the context of Indian art music, Ranjani and Sreenivas (2013) carried out a hierarchical classification of concert sections in the South Indian genre of Carnatic music based on rhythmicity and percussiveness, strongly signaled by the onsets and low frequency content of the accompanying percussion instrument. Thoshkahna et al. (2015) exploited the salience of the estimated tempo to distinguish sections with ambiguous tempo (*alapana*) from the later concerts sections with clear rhythmic properties in Carnatic music concerts. Rhythmic analysis of Indian and Turkish music was explored by Srinivasamurthy et al. (2014), where beat tracking, meter estimation and downbeat detection were identified as musically relevant tasks that could benefit from computational methods. The rhythmic description of Indian music must consider multiple time spans at various levels of hierarchy. Gulati and Rao (2010) explored the use of different signal processing methods for rhythm pattern extraction and evaluated these for tempo detection in North Indian classical music.

Segmentation of Indian instrumental alaps into alap-jod-jhala sections based solely on the rhythmic attributes viz., tempo and its salience, was proposed by Verma et al. (2015), and homogeneity within alap sections was enhanced with the use of posterior probability features from unsupervised modeling. Recognizing that the instrumental music has structural similarities with Dhrupad vocal music, the rhythm features and unsupervised segmentation developed for the plucked strings were tested as such on a small vocal alap dataset. The observed low segmentation performance in this case was speculated to come from the inherently harder problem of syllable onset detection in vocals, and future work suggested "to explore new features" for Dhrupad vocal concert segmentation.

In the present work, we start with rhythmic cues, given the most obvious musical property that distinguishes alap sections, namely the tempo. Motivated by the observations of Verma et al. (2015), we explore new features based on other known and observed musical properties to improve the robustness of vocal alap segmentation over a larger and more structurally diverse dataset. In particular, the dataset now includes concerts with a varying number of sections necessitating new work to handle this unknown. The self-distance matrix provides an excellent framework for visualization, and subsequent computation of the section boundaries in a completely unsupervised manner. Given that the SDM with a suitable choice of feature vector has been an influential method of structural segmentation, we use it as a baseline system in this work. We further explore two supervised methods using the newly enhanced set of melodic-rhythmic features for the detection of structural boundaries, viz. a random forest classifier and a CNN classifier.

## 2. Data Annotation and Evaluation

A dataset for the structural analyses of Dhrupad vocal concerts was put together from available commercial recordings of live performances by leading musicians of the genre. Full length Dhrupad alaps were extracted from concert recordings of five leading artists viz. Gundecha Brothers, Uday Bhawalkar, Ritwik Sanyal, S Wasifuddin Dagar and Sulabha – Manoj Saraf. Nearly half the audios in the dataset are duet performances by two male musicians, the Gundecha Brothers, and one is a duet by a male and a female vocalist. In the duet performances, while the musicians sing alternately most of the time, some phrases are sung in unison. All the musicians are senior exponents of Dhrupad's prominent Dagar tradition, having performed since the mid 1980s.[1,2] The performances show structural diversity, with the number of sections ranging from 3 to 6, with some containing additional faster versions of jod and jhala. The concerts are rendered in different ragas. The concert audios were annotated for the major structural boundaries. We finally have a dataset comprising 20 alap audios of a total duration of 762 minutes, with 53 section boundaries as presented in Table 8, with Section 7 providing a link to further information in the interest of reproducibility.

### *2.1 Data Characteristics and Annotation*

The Dhrupad alap can be viewed as a succession of states, each manifested in an audio segment with a certain musical 'role' with a characteristic behaviour (Peeters and Deruty, 2009). The presence of regular pulsation and its speed are the most distinct properties of a section within a concert. The section boundary itself is cued in multiple ways with both static and dynamic behaviours as presented in **Table 1**. Towards the end of a section, the vocal melody has typically progressed to the highest notes in the range. The melodic ornament, *noom*, is executed next, ending at the middle tonic before a mohra is uttered signifying a change. Given the multiplicity of musical cues, manual labeling of a boundary is expected to have a strong element of subjectivity. We used a consensus based approach involving a discussion between one of the authors and a trained Dhrupad artist. Given that the pulsation speed associated with a section is the most immediately recognised property for a listener, the boundary between sections was marked consistently at the onset of the vocal phrase that introduced the new syllable rate, typically the mohra. This ensured the repeatability of the labeling. However, the acoustic features computed in this work correspond to the distinct melodic and rhythmic cues of **Table 1**, which are actually spread out in time to varying extents (between 3 s and 20 s) in the different concerts.

**Figure 2**, computed with speech analysis software PRAAT (Boersma and Weenink, 2017), shows an excerpt with the glide *noom* and the mohra appearing at the jod-jhala transition. In this case, the mohra is the first vocal phrase at the higher speed of the jhala. It must be noted that the boundary cues, *noom* and mohra, can occur independently within concert sections as well, and when they occur at the boundary between sections, they can be separated by varying extents of silent pauses in the singing. This temporal spread of the various cues indicates the ambiguity inherent in the musical boundary instant location.

The concert sections are of unequal duration, of the order of several minutes, and vary considerably across concerts. Some alaps contain a faster (jalad) section of jhala, while a few contain a jalad-jod section. **Figure 3** displays the diversity of section durations and mean tempi across concert sections in the dataset. We note that,

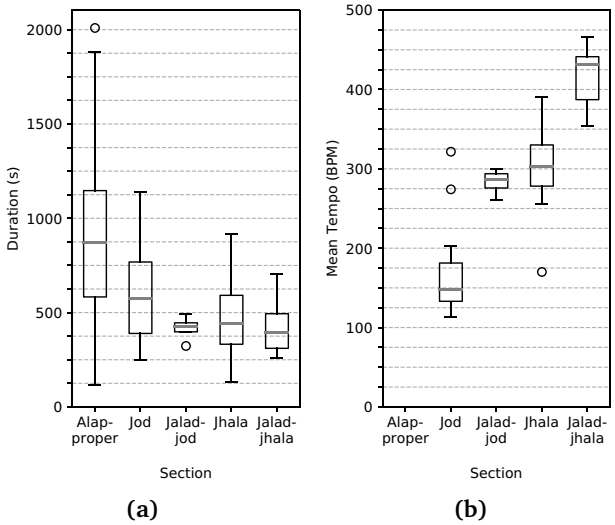


**Figure 3:** Distribution of **(a)** section durations and **(b)** mean tempi of Dhrupad alap sections.

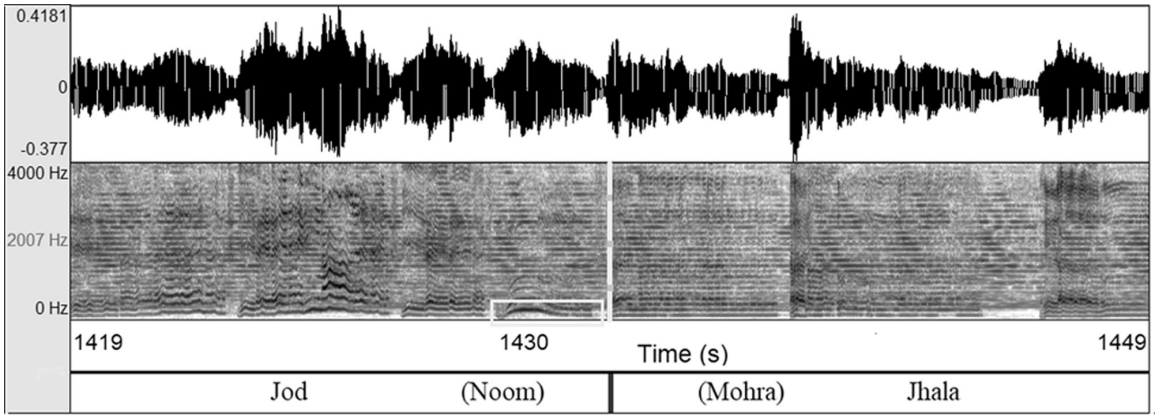


**Figure 2:** Waveform and spectrogram of a 30 s excerpt around a Jod-Jhala boundary (labeled with a vertical dashed line) showing the melodic glide *noom* (with its first harmonic in the box).

unlike the case of the Western music song, a section lasts for several minutes. The time scale of variations is thus relatively large and expected to influence the choice of feature analysis contexts and parameters.

In alaps containing more than three sections, the sections are labeled as either jhala or jod based on the regularity of syllable articulation and the melodic range spanned (Clayton, 2001). We see that the later sections have higher tempi while there is some overlap across the different concerts.

### *2.2 Train-Test Sets and Evaluation Criteria*

The structural segmentation of vocal alap is implemented in this work via the automatic detection of the major section boundaries. We have 20 concert recordings with 53 ground truth boundaries marked in our dataset, described by Table 8. In the interest of testing over a large enough dataset while avoiding train-test leak, we evaluate our supervised segmentation systems using leave-one-concert-out 20-fold cross validation as depicted in **Figure 4**. In each fold, the 19 concerts forming the training set are time- and pitch-shifted in order to obtain an augmented train set. The audios are time-shifted by delays of 0.1, 0.2, 0.3 and 0.4 seconds and pitch-shifted by –2, –1, 1 and 2 semitones.[3] Although the time shifting does not really alter the signal, it changes the frame-level acoustic feature values. The (held-out) test concert is subjected only to time-shifting in order to maintain the acoustics of the test data as such. With 5 test concerts per fold, we finally report performance that is averaged over the augmented test set of 100 concerts in each of the supervised and unsupervised system evaluations.

Viewed as a concert section boundary detection task, we examine uniformly spaced intervals of the audio signal (of duration 1 s, as explained in next section) for the presence or absence of a boundary. A detected boundary is declared a hit if the prediction is within ± *Tol* seconds of a ground truth (GT) boundary; otherwise, it is a false positive. Based on the average inter-judge labeling differences noted by Verma et al. (2015), and also consistent with the observed temporal spread of the different musical cues discussed in the previous section, we report performance with an evaluation tolerance of 15 s using the usual measures of precision, recall and F-value.

## 3. Acoustic Characteristics and Features

In Dhrupad alap, the voice is accompanied only by the drone which tends to have a nearly flat and static harmonic spectrum. Of the musical properties presented in **Table 1**, the perceptually most salient characteristic of a section is the rhythm in the form of the local syllable-rate or tempo. Further, the melodic development across a section is gradual with a reset occurring at the boundary between sections. Motivated by these observations, we consider multiple acoustic features as described next and summarised by the feature extraction flow-chart of **Figure 9**.

### *3.1 Rhythm Features*

The basic dimensions of Indian classical music are melody and rhythm but at the largest time scales relevant to concert structure, rhythm is the prominent distinguishing attribute (Clayton, 2001, p. 96). In a broader sense, rhythm refers to all the aspects of musical time patterns such as the way syllables of the lyrics are uttered, the way the strokes of a musical instrument are played or the inherent tempo of a melodic piece. A rhythm representation can be derived by observing the regularity of note event onsets over a suitably long duration.

#### 3.1.1 Vocal Onset Detection

The rhythm in Dhrupad alaps arises from the rhythmic rendering of vocal syllables. The onset of a syllable is marked by a transient event, characterized by a sudden burst of energy or a change in the short-time spectrum of the audio signal. A computationally simple and effective method of note onset detection involves calculating the temporal derivative of the short-time energy (Bello et al., 2005). The syllables typically uttered – *na*, *re*, *ti*, *de*, are marked by a prominent energy rise in the frequency band of 600–2800 Hz at the consonant-vowel transition (Kumar et al., 2007). The sub-band energy at frame $n$ is given by Equation 1,

$$SB\_Ener[n] = \sum_{k} |X[n,k]W[k]|^2 \tag{1}$$

where $k$ is the frequency bin index, $|X[n, k]|$ is the spectral amplitude feature computed using the short-time DFT of the input audio signal and $W[k]$ is the band limiting filter response with unity gain in the 600–2800 Hz band. The short-time spectrum is computed with a sliding hanning window of duration 30 ms and a hop of 10 ms corresponding to a frame rate of 100/s. An onset instant is then a peak in the derivative of $SB\_Ener[n]$. A robust estimate of the derivative is obtained by incorporating some smoothing prior to differencing via a bi-phasic

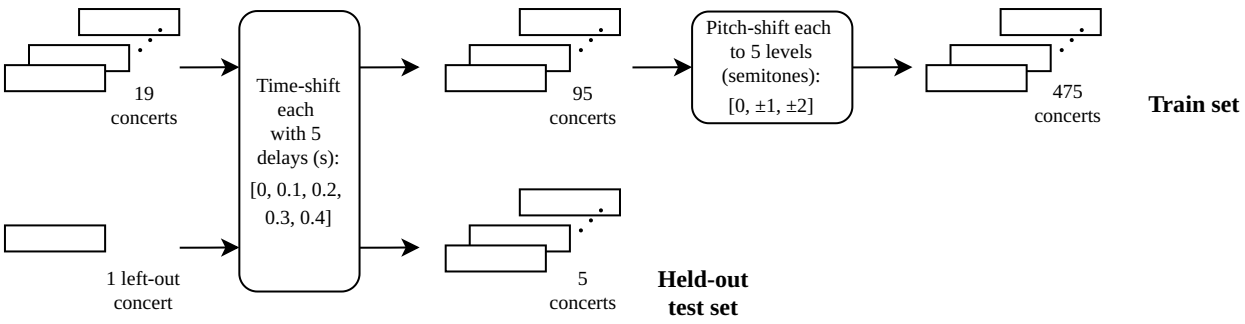


**Figure 4:** One instance (fold) of the 20-fold cross-validation process adopted for train-test data splitting.

function serving as a filter. A discrete time filter, $h[n]$, is obtained by sampling the impulse response of the biphasic filter recommended for vowel onset detection (and given by Eq. 4 in (Hermes, 1990)) at the required 10 ms frame intervals. **Figure 5a** shows a plot of $h[n]$ superposed on the underlying continuous-time biphasic function whose parameters are the lobe widths and locations of its two peaks. The same filter was used effectively in the context of sung and hummed notes by Kumar et al. (2007). An onset detection function (ODF) is obtained then by the convolution of the sub-band energy function with the filter impulse response.

$$ODF_{syll}[n] = SB\_Ener[n] * h[n] \quad (2)$$

The quality of the onset detection function is expected to influence the reliability of the rhythm representation derived from it. We therefore tested the ODF independently on a selected diverse set of 130 labeled vocal syllable onsets across 6 concert segments spanning jod and jhala sections. We obtained a recall of 0.7 and a precision of 0.8 at the peak-picking threshold corresponding to the best F-score of 0.75. The performance was observed to be superior in jhala due to the more regularly articulated syllables relative to the jod utterances. However, the onset detection performance overall on vocal syllables is significantly lower than that achieved on sitar and sarod plucks in instrumental concerts (Vinutha et al., 2016). This attests the challenge in vocal music posed by the greater diversity of phonetic realisations and singing styles.

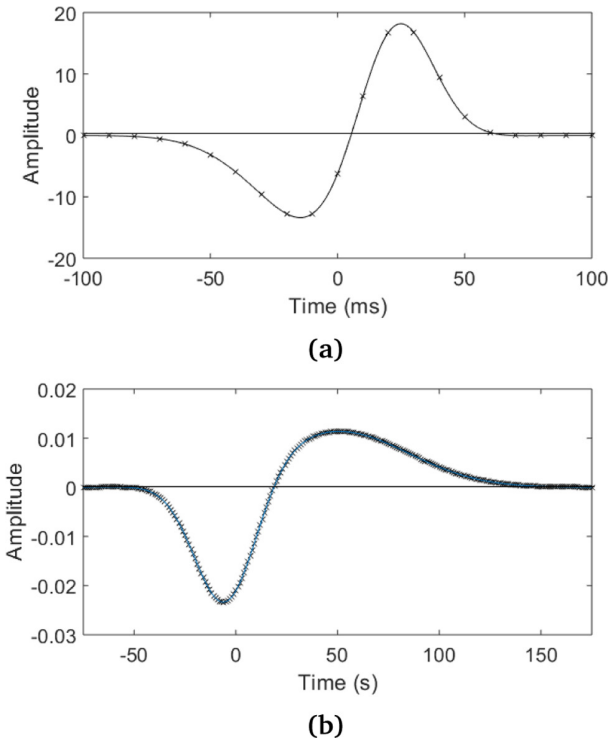


**Figure 5:** Bi-phasic filter impulse response, with the discrete samples superposed, as applied to generate **(a)** onset detection function from sub-band energy, and **(b)** derivative features from short-time energy and short-time spectral centroid.

### 3.1.2 Tempo Estimation

The local tempo can be estimated by measuring the periodicity of the onset detection function. Given the frequent occurrences of brief, intermittent silences in the singing, we choose a window duration of 20 s over which to compute the short-time autocorrelation function (ACF). The ACF of the onset detection function (sampled at 10 ms intervals) is computed for up to 300 lags (range of 0–3 s, which spans several tens of pulses at the lowest expected tempo of 100 BPM). A powerful visualization of the periodicity captured by the short-time ACF is seen in the image of ACF strength versus time and lag in **Figure 6** known as a rhythmogram (Jensen et al., 2005). We observe the absence of periodic structure in the alap-proper section, while the jod and jhala sections are characterized by a strong periodicity, indicated by the horizontal striations equispaced in lag. The decreasing separation between striations indicates an increasing rate of onsets, i.e., increasing tempo. The boundaries between the sections are clearly visible in the rhythmogram, suggesting that the ACF could serve as a feature vector for SDM-based segmentation.

The ACF is a high dimensional vector that embeds the rate of pulsation given the absence of metrical hierarchy in the Dhrupad alap. It can potentially be replaced by a single tempo value. A reliable method of tempo detection combines the ACF and DFT in a product that yields tempo estimates relatively free from octave error (Peeters, 2007). Next, the normalized ACF peak value corresponding to the detected tempo is used as a measure of salience or pulse clarity, and detected tempo values with a low salience (<0.1) are clamped to zero. We thus obtain the two dimensional vector (tempo, salience) as a compact alternative to the high-dimensional ACF vector. **Figure 7a** and **7b** show the time-varying tempo and salience respectively. We observe that within the jod and jhala sections the tempo gradually increases, with a jump at the boundaries, while in the rhythm-free alap section the estimated salience is uniformly low and detected tempo, random.

### 3.1.3 Posterior Features

Verma et al. (2015) showed that transforming the rhythm feature vector of (tempo, salience) to a vector of class-conditional probabilities (or posteriors), where the classes

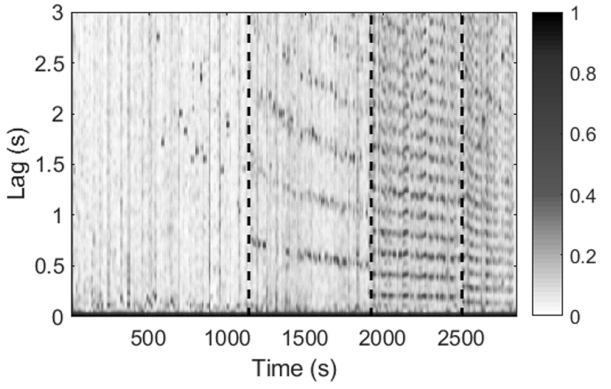


**Figure 6:** Rhythmogram of the UB_AhirBhrv alap (dashed lines indicate labeled boundaries).

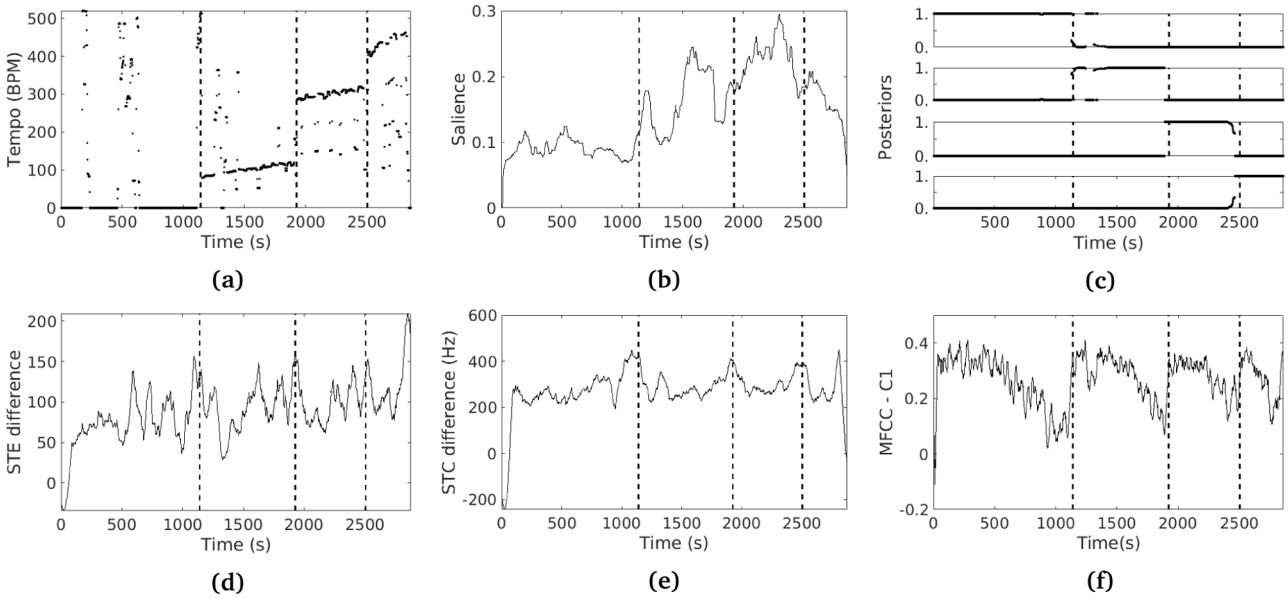


**Figure 7:** Analysis of UB_AhirBhrv alap containing 4 sections: alap-proper, jod, jhala and jalad-jhala **(a)** Tempo, **(b)** Salience or pulse clarity, **(c)** Posteriors of rhythm, **(d)** Short-time energy difference, **(e)** Short-time centroid difference, and **(f)** MFCC C-1 coefficient. Dashed lines indicate manually labeled section boundaries.

comprise the distinct sections, improves homogeneity of the resultant features within a section. Each feature vector $V_i$ of a frame $i$ of a concert is transformed to a vector $q_i$ whose length matches the estimated number of sections, $K$. We derive the posterior features by the unsupervised clustering of the rhythm using a GMM model with $K$ Gaussians ($C_1$, $C_2$, ..., $C_K$) representing the $K$ sections. Thus, each $q_i$ comprises:

$$q_i = \left(P(C_1 \mid V_i), P(C_2 \mid V_i), ..., P(C_K \mid V_i)\right). \quad (3)$$

In Equation 3, the k-th dimension of $q_i$ represents the posterior probability $P$, given the frame vector $V_i$, of the k-th Gaussian component. The GMM is trained with maximum likelihood across all the frames in a given concert (i.e. in an unsupervised manner). We expect noisy feature values, which in turn affect the homogeneity of a section, to be ideally mapped to low probability values in the posterior vector. **Figure 7c** plots the posterior probabilities with time. We see that the unsupervised clustering has indeed resulted in peaky posteriors with only one or the other of the 4 probabilities in the posterior vector (presumably the one corresponding to the Gaussian representing that section) dominating in a given ground truth section, with relatively sharp transitions at the boundaries.

Verma et al. (2015) set the number of sections to a fixed value (K = 3) since this was the ground truth across all 10 concerts in their dataset. Given the greater diversity in the current dataset (where the number of sections ranges between 3 and 6), the question arises of estimating the number of sections, K. We apply the Bayesian Information Criterion (BIC), a likelihood criterion penalised by the model complexity, to estimate K in the expected range (Chen and Gopalakrishnan, 1998). We choose the value of K that minimises the BIC criterion for the given concert. This gives us finally a variable dimension vector of posterior probabilities. It was observed that the number of sections is estimated correctly in 13 of the 20 concerts and over-estimated by 1 or 2 in the remaining, typically longer duration, concerts.

### *3.2 Melody and Timbre Features*

We note from Section 2.1, that a concert section is marked by the progression of the melody from the middle octave, down, and then up to the higher reaches. We see this in **Figure 2** where the pitch (as indicated by the harmonic spacing) is consistently higher before the boundary relative to that just after. The transition between sections is therefore marked by a prominent reset in the singing pitch. Pitch and chroma features essentially manifest this change but are found to be affected similarly by the melodic development within the section apart from the challenges presented by the nearly 3-octave range. A related but more robust attribute is found in the changing loudness and brightness of the voice with pitch, arising from the increased sub-glottal pressure or vocal effort required to produce higher pitches (Sundberg, 1990). We therefore examine the use of timbre features in section boundary detection. The short-time log magnitude spectrum computed on the auditory mel scale (log mel-spectrogram), using 40 filters in the 80 Hz–8 kHz band, is shown in **Figure 8** across duration for the UB_AhirBhrv alap. We can see the shift in energy away from the lower frequency bands as the melody attains its height near the boundary. This is accompanied by an increase in frequency spreading. This changing spectral shape, due to changing vocal intensity and brightness, can be compactly represented by the short-time energy and the spectral centroid respectively. We consider these in our feature design as also mel-frequency cepstral coefficients (MFCC), given their ubiquity in music classification tasks involving timbre (Logan, 2000). We include the 13-dimensional MFCC vector (coefficients C-0 to C-12) in our set of features for evaluation.

The above features are computed at the 10 ms frame level, over 30 ms long sliding windows, and then subsampled to a 1 s frame level after averaging over suitably long sliding windows. The length of the averaging window controls the smoothing and a longer window helps remove the fine fluctuations irrelevant at the larger timescales that we are interested in. We experiment with two relatively extreme values for the averaging window length – a short 3 s window relating to the duration of the *noom* glide, and a much longer one, 20 s, relating to longer-term trends in melodic pitch. With the intention of obtaining the section boundaries as the instances of change, we further compute derivatives of the short-time energy and spectral centroid features by convolving each with the discrete version of a biphasic filter given by **Figure 5b**. The peak width and location parameters were experimentally tuned to maximize the strength of peaks in the immediate vicinity of labeled boundaries. The outputs of the filter are referred to as the short-time energy (STE) difference and short-time centroid (STC) difference.

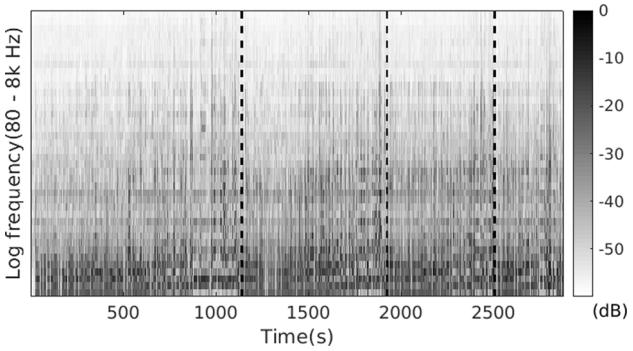


**Figure 8:** Mel-spectrogram of UB_AhirBhrv alap (dashed lines indicate labeled boundaries).

**Figure 7d** shows the frame-level short-time energy difference with the anticipated sharp rise at the boundaries but also several peaks within each section. **Figure 7e** shows the spectral centroid difference, where again, peaks can be seen occurring at the marked ground truth boundaries. **Figure 7f** shows the second MFCC coefficient (C-1), associated with spectral envelope tilt, rising sharply at the start of every section and gradually dropping as the section progresses. We observe that both the spectral shape indicators are characterised by clearer boundary effects compared to the short-time energy.

## 4. Structural Segmentation Methods

In our context, structural segmentation involves the detection of change between contrasting musical parts dictated by one or more musical attributes. This can be viewed as a boundary detection task where each frame, occurring at the rate of 1 Hz, is to be classified as a boundary or non-boundary frame based on whether a transition between musical sections occurs over the frame duration. The features, computed as presented in **Figure 9**, and summarised in **Table 2**, are individually

**Table 2:** A key to the naming of feature subsets.

| Feature subset name | Features |
|---|---|
| Rhythm | Posteriors of (tempo, salience) |
| MFCC | First 13 MFCCs |
| Timbre | MFCC<br>Short-time energy difference<br>Short-time centroid difference |
| All | Rhythm<br>Timbre |

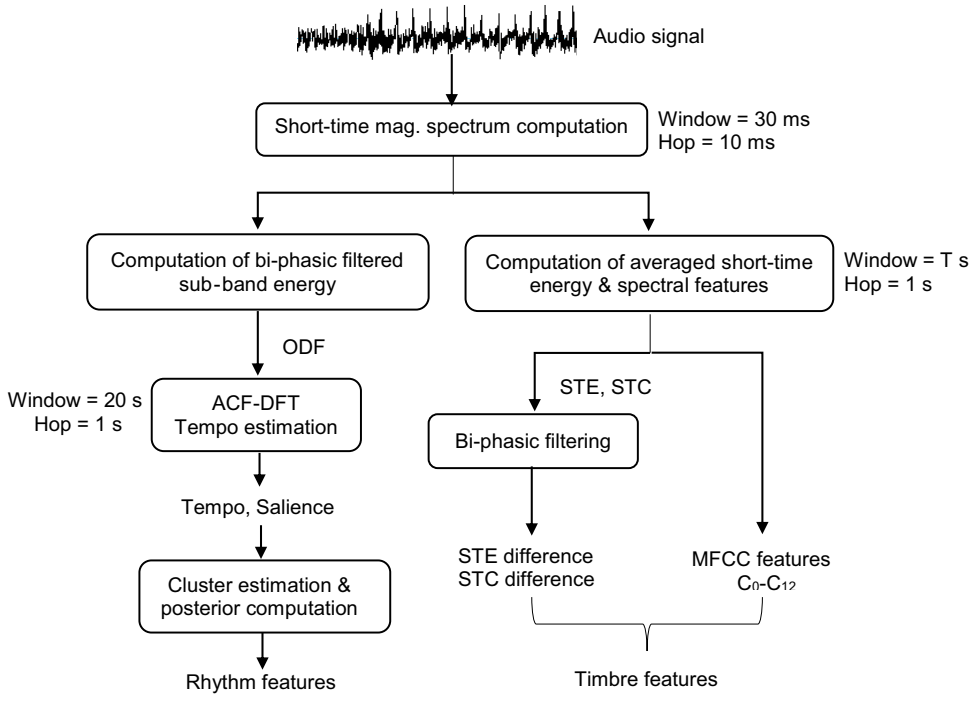


**Figure 9:** Block diagram for the extraction of acoustic features.

z-score normalized across each concert to obtain a mean of 0 and a standard deviation of 1 to derive the classifier inputs. Boundary detection can be achieved by an unsupervised framework involving the SDM and kernel correlation or by the supervised classification of the 1 s frames as boundary/non-boundary events. We also investigate feature learning, from the (relatively low-level) log mel spectrogram representation, via a CNN classifier. In all cases, the evaluation uses the 100 time shifted audios generated from the original 20 concerts to obtain reliable measures of boundary detection performance.

### *4.1 Unsupervised Segmentation*

Given a feature vector stream, the SDM can be computed using a chosen distance measure, in our case the L2 distance (Paulus et al., 2010). Thus, a homogeneous segment of length M frames would appear as an M × M block of low distance values. Next, points of high contrast in the similarity matrix are captured by convolution along the diagonal with a checker-board kernel of dimension matched to the time-scale of interest (Foote, 2000). Given that the minimum section duration is about 100 s in the dataset, we examine kernels of size 50 × 50 and 100 × 100. The one-dimensional plot resulting from the convolution is called a novelty function, whose peaks indicate the boundary time instants in the feature vector stream.

**Figures 10** and **11 (a, b & c)** present the SDM and novelty function respectively for the ACF, rhythmic features and the posterior features for the UB_AhirBhrv alap with the 100 × 100 kernel. The novelty function derived from the ACF is observed to be noisy. The rhythmic features, tempo and salience, improve upon this to an extent, while the posterior features visibly improve the homogeneity of the sections and consequently the accuracy of detected peaks in the novelty function, confirming the observations made by Verma et al. (2015) for instrumental concerts.

It was observed that, rather than combining both rhythm and timbre features into a single vector for the SDM, fusing the information in the distinct feature streams at the peak picking stage provided more flexibility in terms of tuning the performance of the system. The STE-difference and STC-difference already represent feature derivatives and are treated directly as novelty functions for peak picking. The SDM for the 13-dimensional MFCC vector and the corresponding novelty function obtained by convolution with a chosen kernel size (100 × 100) are shown in **Figures 10(d)** and **11(d)** respectively. We observe clear peaks at the labeled boundary instants, but also a few spurious peaks within sections arising from local timbre variations.

Next, the highest *N* (varied from 1 to 18, treated as a tunable parameter) peaks are picked in each feature's novelty function stream as boundary predictions, while ensuring that no two selected peaks are within 30 s of each other. For the information fusion, MFCC is taken as the reference as it is found to perform best among individual feature categories. Boundary candidates derived from novelty functions of the rhythm feature vector, MFCC vector and each of the two 1D timbre features are fused

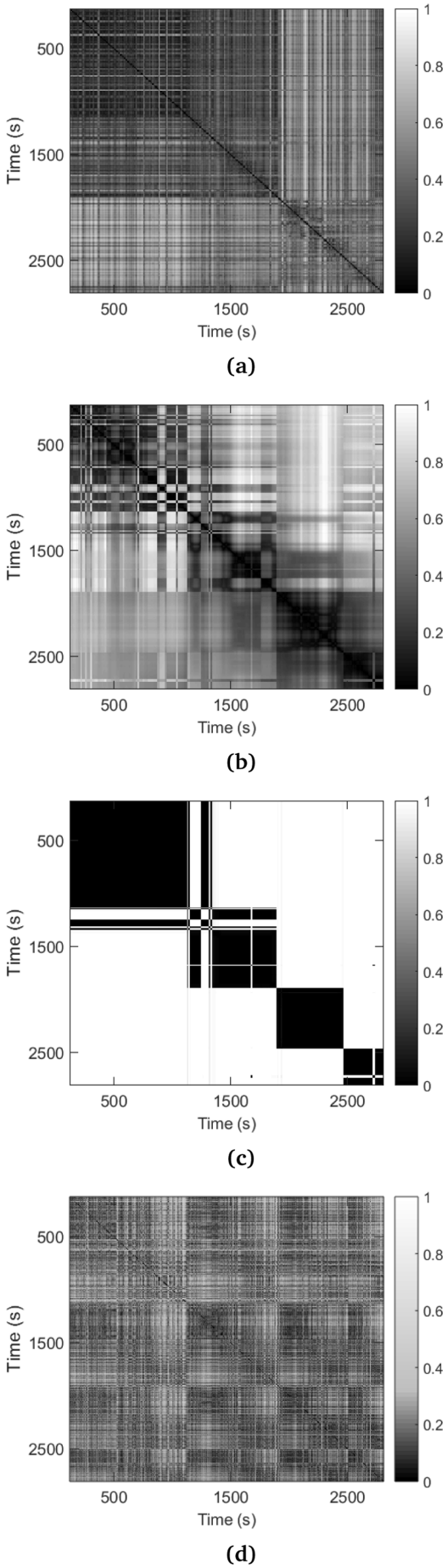


**Figure 10:** SDM for UB_AhirBhrv alap with **(a)** ACF, **(b)** rhythm features, **(c)** posterior features, and **(d)** MFCC.

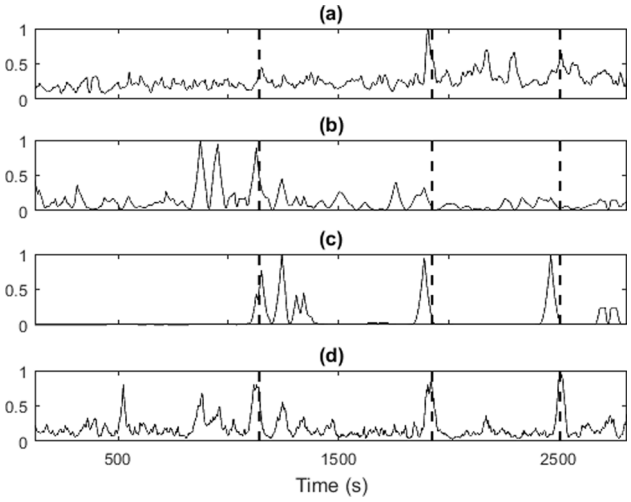


**Figure 11:** Novelty function for UB_AhirBhrv alap with **(a)** ACF, **(b)** rhythm features, **(c)** posterior features, and **(d)** MFCC. Dashed lines indicate manual boundaries.

using a majority rule (i.e. two or more features out of three are checked for coincidence with the MFCC reference).

### *4.2 Supervised Segmentation Methods*

As mentioned in Section 2.2, classifiers are trained on the augmented dataset that includes the pitch and time shifted versions of all the audios in each fold. To account for ambiguity in manual boundary marking from both, the variety and temporal spreading of the section change cues, targets are smeared by labelling all the frames in a ±15 second window about the manually labeled boundary as "boundary" frames (Ullrich et al., 2014). Next, in order to balance the non-boundary and boundary frame examples in training, only the boundary-labeled frames of the newly generated audios are retained, along with all of the frames of the original dataset.

Based on the expectation that frame-level boundary detection would benefit from context, features from adjacent frames in a $\pm C$ s neighbourhood are appended to the current frame features. The corresponding target is a label indicating the presence or absence of a manually labeled boundary at the center frame. The classifier output is the estimated probability of a boundary in the frame. During evaluation, the frame-wise boundary predictions are post-processed by replacing predictions within a 30 s window with a single one of the highest strength (probability) given that the distinct acoustic cues to a section boundary can be spread over several seconds. The obtained frame-wise values are compared with a threshold to obtain the detected boundaries. The threshold is varied to obtain a Receiver Operating Characteristic (ROC) curve and to choose the point of best performance in terms of F-score with reference to the manually labeled boundaries.

#### 4.2.1 Random Forest Classifier

A random forest classifier consists of an ensemble of decision trees and outputs the class with the majority vote (weighted by the corresponding probability values) as the model's prediction. The classifier is trained on input vectors of features (posterior rhythm and timbre) including those of the current and context frames. A target of 1 or 0 is assigned to each input training vector indicating whether the current frame is a manually labeled boundary or not. The posterior rhythm feature is a vector whose length is ideally equal to the number of sections in the alap, meaning that the length is not the same for every concert. Moreover, the values of this feature at every frame indicate the probability of the frame belonging to a particular section, and hence, for all the frames within a section, only one of the posteriors dominates, whereas very near a boundary, the values are observed to be more distributed in the [0,1] range due to the non-homogeneity of frames close to the boundary. Therefore, only the maximum value in the posterior rhythm vector is used as the corresponding feature.

While there are several model hyperparameters in a random forest classifier that can be tuned to optimize the performance, we focus mainly on optimizing the number of decision trees, and leave the others at their recommended values.[4] In addition, we experiment with some hyperparameters related to feature extraction, such as averaging window size and context duration. The number of decision trees is varied between 10 and 100 in steps of 10, while the context duration ($C$) is swept from ±10 to ±50 seconds in steps of 10 s. The window used to average the timbre features over is set to each of the two values, 3 s and 20 s, while the window for the rhythm feature computation from the onset detection function is always set to 20 s, as explained earlier.

#### 4.2.2 Convolutional Neural Network

We investigate the direct learning of features from the data by the use of a convolutional neural network (CNN)-based classifier. A relatively generic form of CNN was successfully employed for music structure analysis on the large SALAMI dataset (Ullrich et al., 2014). In our work, the features, or representation, are learned from the input mel-spectrogram (as described in Section 3.2) by training with frame-level targets indicating the manually labeled boundary and non-boundary frames. As we wish to obtain boundary predictions at a 1 s frame resolution, the log-mel-spectrogram is then sub-sampled by averaging within overlapping windows of a suitable size, with a hop of 1 s (in line with the use of averaging windows in the feature extraction step described earlier). The mel-spectrogram is next split into smaller overlapping chunks of size $40 \times N_C$, by taking $\pm N_C/2$ adjacent frames as context (corresponding to $\pm C$ s) for each frame input to the network. The input to the network is normalised so that all values in a single input chunk lie in the range −1 to +1. The target output is 1 or 0 indicating a manually labeled boundary or none in the center frame.

The CNN model architecture used in this work is based on Ullrich et al. (2014) with two conv. layers, the first with 16 filters of 6 × 6 kernels (6 time frames, 6 mel-bands), and the second with 32 filters of 3 × 3 kernels. Each conv. layer is followed by a max-pool layer of dimension 3 × 3. Next, we use a fully-connected (FC) layer with 128 hidden units and finally an FC layer (output) with 2 units. Each of the conv. layers applies a

ReLU activation, the first FC layer applies a sigmoid, and the output layer applies a softmax activation. We also experimented with non-square kernels and with other combinations of activation functions – using the tanh instead of ReLU for the conv. layers, and using ReLU for all the layers. These initial experiments are used to fix the model hyperparameters, before trying to optimize the input feature hyperparameters like the averaging and context window durations. We experiment with two values for the averaging window (duration over which the mel-spectrogram is locally averaged with a sliding window) – 3 s and 20 s, and in each case vary the context duration ($C$) from ±10 to ±50 s in steps of 10 s.

The model is trained using binary cross-entropy loss on mini-batches of 64 samples, over 20 epochs, using an Adam optimizer with an initial learning rate of 0.001. The model is trained and evaluated using a leave-one-concert-out method as depicted in **Figure 4**. The epoch resulting in the least validation loss computed on the held-out concert is used to obtain the boundary predictions on the corresponding concert.

## 5. Results and Discussion

We present the segment boundary detection performances of our baseline unsupervised system and the two supervised (RF and CNN) methods as averages obtained across our test set of 100 concerts including the time-shifted versions of the original audios. We report results for the distinct, as well as combined, feature subsets specified in **Table 2**.

### *5.1 Unsupervised Boundary Detection*

**Table 3** presents the unsupervised boundary detection performance obtained using the different feature sets. Testing across the system hyperparameters for the rhythm and timbre feature categories, a shorter averaging window of 3 s for the STE and STC features, and a wider 100 × 100 checker-board kernel for the SDM (used for the rhythm and MFCC feature vectors) are found to result in higher F-scores relative to the other considered alternatives. The optimal number of candidate peaks for each novelty function stream is found to be N = 7. It can be seen that while the timbre features bring only a small improvement over the sole use of MFCC, the fusion of timbre and rhythm features is clearly superior to segmentation based on timbre alone.

### *5.2 Random Forest Classifier*

**Table 4** shows the best boundary detection performance obtained using each feature subset, and the corresponding values of the hyperparameters. An averaging window duration of 20 s for the timbre features was found to yield the best performance and hence all the results are reported with this setting. It can be seen that the rhythm features alone do not perform well, while the timbre features do significantly better, and a combination of the rhythm and timbre features results in a small further boost to the performance.

It is interesting to note that MFCCs alone perform comparatively well, but with a much higher context duration than when using all the timbre features. However, the improved score using all the timbre features comes at a cost of a higher number of trees. Adding rhythm features increases the F-score further only slightly, and at the cost of even more trees in the classifier. Further, the best F-score for each feature subset is not obtained necessarily at the highest values of the context duration and number of trees, suggesting that after a point the model starts to overfit. Also interesting, although not explainable, is that while in the unsupervised case a 3 s timbre averaging window worked best, the RF classifier did best with a 20 s window.

We also train the model without any of the data augmentation discussed in Section 2.2 to assess its contribution to the overall performance. These results are reported only for the all features condition, appearing within parentheses in **Table 4**. It is seen that augmentation clearly helps improve the recall in boundary detection.

**Table 3:** Performance of the unsupervised approach using different feature subsets with specified averaging window durations.

| Feature subset | Parameters | Performance | | |
|---|---|---|---|---|
| | Window(s) | Precision | Recall | F-score |
| Rhythm | 20 | 0.40 | 0.57 | 0.47 |
| MFCC | 3 | 0.59 | 0.57 | 0.58 |
| Timbre | 3 | 0.61 | 0.57 | 0.59 |
| All | 20 or 3 | 0.72 | 0.66 | 0.69 |

**Table 4:** Performance of RF classifier using different feature subsets with an averaging window of 20 s, for values of context ($C$) and #trees giving the best F-scores. In parentheses are results without training data augmentation.

| Feature subset | Parameters | | Performance | | |
|---|---|---|---|---|---|
| | $C$ (±s) | # trees | Precision | Recall | F-score |
| Rhythm | 50 | 30 | 0.17 | 0.26 | 0.21 |
| MFCC | 50 | 10 | 0.85 | 0.74 | 0.79 |
| Timbre | 20 | 50 | 0.86 | 0.81 | 0.83 |
| All | 20 | 100 | 0.90 (0.89) | 0.81 (0.75) | 0.85 (0.81) |

### *5.3 CNN*

In the CNN model-related experiments we started from the architecture described in Section 4.2.2, and experimented with rectangular kernels for the conv. as well as the pool layers. The rectangular kernels in the pool layer were made longer only in the frequency dimension in order to preserve temporal resolution. During these preliminary experiments, the averaging window and context duration were set to 3 s and ±20 s, respectively. The final architecture that yielded the best results had kernels of size 3 × 6 in the conv. layers (3 along frequency and 6 along time), and of size 3 × 3 in the pool layers.

**Table 5** shows the results obtained using this architecture, for two context durations, ±20 s and ±50 s (motivated by the results from the RF classifier). Results are reported only for the case with an averaging window duration of 3 s, since these were significantly better than with a longer 20 s. It is evident from the results that the CNN classifier performs better with a larger context duration of ±50 s which provides a wide view of the pre- and post-boundary acoustic cues. Changes to the activation functions in the conv. and the fully-connected layer did not affect the F-score.

### *5.4 Discussion*

The boundary detection performances of all the three approaches are consolidated in **Table 6** for the best hyperparameter settings of each as determined in the previous sections. Given that the CNN classifier does not receive an explicit rhythm representation of the signal, we present in addition the other methods minus rhythm features. The supervised approaches, CNN and RF classifier, perform equally well while the unsupervised approach is clearly worse. With the low-dimensional rhythm features alone, the unsupervised system performs better than the supervised RF classifier (as we saw in **Tables 3** and **4**). The latter benefits only slightly in precision from the rhythm features. An interesting observation was that the novelty peaks from the rhythm features signaled the boundaries more distinctly than did the timbre features in certain instances where the latter were unreliable such as the male-female duet where the melody reset at the section boundary was obscured by the constant switching between the individual singers' ranges. Across concerts, false positives were seen when boundary-like cues such as the noom and mohra appeared within sections (i.e. without a tempo increase), signaling here changes at some lower level in the structural hierarchy. Missed detections were observed in instances where the tempo change between sections was relatively low. The interaction of multiple cues is evident overall.

**Table 5:** Performance of CNN classifier with averaging window of 3 s with different context durations C.

| Parameters | Performance | | |
|---|---|---|---|
| *C* (±s) | Precision | Recall | F-score |
| 20 | 0.69 | 0.77 | 0.73 |
| 50 | 0.92 | 0.81 | 0.86 |

**Table 6:** Performance comparison of different feature combinations and methods.

| Segmentation approach | Precision | Recall | F-score |
|---|---|---|---|
| **Without rhythm features** | | | |
| RF | 0.86 | 0.81 | 0.83 |
| CNN | 0.92 | 0.81 | 0.86 |
| Unsupervised | 0.61 | 0.57 | 0.59 |
| **With rhythm features** | | | |
| RF | 0.90 | 0.81 | 0.85 |
| Unsupervised | 0.72 | 0.66 | 0.69 |

We report next the performance of the above configured systems on two new concerts (outside the previous dataset and therefore not included in the systems' hyperparameter optimization). The concerts are further distinguished in that they are by a young, female Dhrupad vocalist,[5] over 20 years apart from our dataset artists. The alaps, each about 20 minutes in duration, contain two boundaries each, demarcating three sections. In a departure from the vocal style of the previous artists, Pelva intersperses the normally uttered Dhrupad syllables with the vowel 'a', thus reducing the clarity of the computed rhythmogram even though the tempo change at section transitions remains perceptually salient. She also sometimes omits the section change cue *noom*. The characteristic shifting of melodic focus across the section and its abrupt reset to the middle tonic are maintained but with one case of a larger than usual separation in time between the tempo cue and melodic cue instants in the second concert. Finally, the recordings are marked by relatively loud tambura (drone) background.

**Table 7** presents the results obtained on an augmented test set of 5 time-shifted audios corresponding to each concert. We observe that the RF classifier performs best, with performance similar to that obtained on the 20 concert dataset indicating that the system generalizes well. This is not true for the CNN classifier where the performance drops steeply, especially in the case of the second concert, affected probably by the above noted concert specific variation. The unsupervised system exhibits a similar performance across the two concerts although the F-score is slightly worse than that obtained on the previous 20 concert dataset.

## 6. Conclusion

The Dhrupad alap is a highly structured performance within an improvisational framework. Rhythm or tempo marks the evolution of the concert in time with abrupt

**Table 7:** Segmentation performance of the different methods on test concerts.

| Test alap | Unsupervised | | | RF Classifier | | | CNN | | |
|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | F-score | Precision | Recall | F-score | Precision | Recall | F-score |
| PN_Jog | 0.50 | 0.50 | 0.50 | 0.90 | 0.90 | 0.90 | 0.50 | 1.0 | 0.67 |
| PN_Maru | 0.50 | 0.50 | 0.50 | 0.47 | 0.70 | 0.56 | 0 | 0 | 0 |

changes at section boundaries. Within each section, melodic development plays out in a similar way with the gradual shifting of melodic focus starting from the concert tonic. The above musical cues were found to be effectively captured with acoustic features related to syllable rate and vocal brightness, both computable from the short-time magnitude spectrum representation of the audio recording. Thus, our work provides us with explicit descriptions about the audible structure of the alap, an important constituent of the listener's unconscious schematic expectations (Widdess, 2011).

With the given training dataset, a supervised classifier trained on the hand-crafted features performed best overall. While the perceptually most distinguishing characteristic of the concert sections is the syllable rate or tempo, the more powerful cues in the automatic detection of boundaries were found to be the abrupt melodic transitions. More reliable means of detection of the vocal syllabic onsets can potentially lead to more robust rhythm features. A take-home message is therefore that it can be rewarding to investigate MIR methods on new dataset/task scenarios, both from achieving reasonable performance outcomes and for drawing a deeper understanding of genre characteristics from the acoustical analyses. While the CNN classifier performs competitively on the 20 concert dataset through a purely learned representation, it is more affected by concert-dependent variations that could have resolved with larger and more diverse training data. The possibility of learning features for structural boundary detection in an unsupervised manner is also an attractive prospect provided a sufficiently large genre-specific dataset can be assembled (McCallum, 2019). Finally, applying the outcomes of this research to the concert summarization task where the important musicological cues to section character are preserved is an interesting topic for future work, together with its extension to the composition sections of the Dhrupad concert that are rendered after the alap (Ranganathan, 2013).

## 7. Reproducibility

All annotations, code and trained models are available at this link: https://github.com/DAP-Lab/dhrupad-alap-segmentation. The annotations contain section boundaries and labels for all the concerts used in the cross-validation and test experiments. The concert audios are not made available, but links to their sources are provided.

## Appendix

**Table 8:** The Dhrupad alap dataset used in this work (see Github repository mentioned in Section 7 for details).

| Sl.# | Alap | Artist | Raga | Dur (min) | #Sections |
|---|---|---|---|---|---|
| 1 | GB_AhirBhrv | Gundecha Brothers | Ahir Bhairav | 49:47 | 4 |
| 2 | GB_Bhg | Gundecha Brothers | Bihag | 21:23 | 3 |
| 3 | GB_Bhim | Gundecha Brothers | Bhimpalasi | 17:22 | 3 |
| 4 | GB_Bhrv | Gundecha Brothers | Bhairav | 53:11 | 6 |
| 5 | GB_BKT | Gundecha Brothers | Bilaskhani Todi | 43:00 | 4 |
| 6 | GB_KRA | Gundecha Brothers | Komal Rishabh Asavari | 36:30 | 4 |
| 7 | GB_Mar | Gundecha Brothers | Marwa | 48:34 | 5 |
| 8 | GB_MMal | Gundecha Brothers | Miya Malhar | 45:42 | 5 |
| 9 | GB_Yam | Gundecha Brothers | Yaman | 46:32 | 4 |
| 10 | RS_Bind | Ritwik Sanyal | Bindeshwari | 19:57 | 4 |
| 11 | RS_Shr | Ritwik Sanyal | Shree | 26:90 | 3 |
| 12 | Sul_Man_Yam | Sulabha – Manoj Saraf | Yaman | 21:46 | 3 |
| 13 | UB_AhirBhrv | Uday Bhawalkar | Ahir Bhairav | 48:00 | 4 |
| 14 | UB_Bhg | Uday Bhawalkar | Bihag | 51:10 | 3 |
| 15 | UB_Bhrv | Uday Bhawalkar | Bhairav | 50:22 | 3 |
| 16 | UB_Jog | Uday Bhawalkar | Jog | 25:46 | 3 |
| 17 | UB_Malk | Uday Bhawalkar | Malkauns | 61:16 | 3 |
| 18 | UB_Maru | Uday Bhawalkar | Maru | 35:35 | 3 |
| 19 | UB_Shr | Uday Bhawalkar | Shree | 19:45 | 3 |
| 20 | WD_Bhg | Wasifuddin Dagar | Bihag | 40:22 | 3 |

## Notes

[1] Dhrupad Kendra, Bhopal — Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/wiki/Dhrupad_Kendra_Bhopal, Accessed: 7 January 2020.

[2] Ritwik Sanyal — Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/wiki/Ritwik_Sanyal, Accessed: 7 January 2020.

[3] Rubber Band Library v1.8.2, https://breakfastquay.com/rubberband/, Accessed: May 2019.

[4] https://scikit-learn.org/stable/modules/ensemble.html#parameters, Accessed: December 2019.

[5] http://pelvanaik.in/about/, Accessed: 7 January 2020.

## Additional File

The additional file for this article can be found as follows:

- **Supplementary file.** Structural Segmentation of Dhrupad Vocal Alaps. DOI: https://doi.org/10.5334/tismir.64.s1

## Acknowledgements

The authors would like to thank Uddalok Sarkar for his contribution to the design of the timbre features. The authors would also like to thank the editor and reviewers for their valuable suggestions.

## Competing Interests

The authors have no competing interests to declare.

## Author Contributions

Preeti Rao contributed to the overall design of the study and the writing of this article. Vinutha contributed to the dataset preparation and the design, implementation and presentation of the unsupervised system. Rohit contributed to the design, implementation and presentation of the supervised system. All authors have read and agreed to the published version of the manuscript.

## References

**Allegraud, P., Bigo, L., Feisthauer, L., Giraud, M., Groult, R., Leguy, E.,** & **Levé, F.** (2019). Learning sonata form structure on Mozart's string quartets. *Transactions of the International Society for Music Information Retrieval*, *2*(1), 82–96. DOI: https://doi.org/10.5334/tismir.27

**Bartsch, M. A.,** & **Wakefield, G. H.** (2005). Audio thumbnailing of popular music using chromabased representations. *IEEE Transactions on Multimedia*, *7*(1), 96–104. DOI: https://doi.org/10.1109/TMM.2004.840597

**Bello, J. P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M.,** & **Sandler, M. B.** (2005). A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, *13*(5), 1035–1047. DOI: https://doi.org/10.1109/TSA.2005.851998

**Boersma, P.,** & **Weenink, D.** (2017). Praat: Doing phonetics by computer [computer program]. Version 6.0.28. http://www.praat.org/. Retrieved March 3, 2017.

**Chen, S.,** & **Gopalakrishnan, P.** (1998). Speaker, environment and channel change detection and clustering via the Bayesian information criterion. In *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, volume 8, pages 127–132, Virginia, USA.

**Clarke, E. F.** (1999). Rhythm and timing in music. In *The Psychology of Music (Second Edition)*, pages 473–500. Elsevier. DOI: https://doi.org/10.1016/B978-0122 13564-4/50014-7

**Clayton, M.** (2001). *Time in Indian Music: Rhythm, Metre, and Form in North Indian Rag Performance*, Chapter 11: A case study in rhythmic analysis. Oxford University Press, UK.

**Cooper, M.,** & **Foote, J.** (2003). Summarizing popular music via structural similarity analysis. In *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 127–130. DOI: https://doi.org/10.1109/ASPAA.2003.1285836

**Dannenberg, R. B.,** & **Goto, M.** (2008). Music structure analysis from acoustic signals. In *Handbook of Signal Processing in Acoustics*, pages 305–331. Springer. DOI: https://doi.org/10.1007/978-0-387-30441-0_21

**Dixon, S.** (2001). Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research*, *30*(1), 39–58. DOI: https://doi.org/10.1076/jnmr.30.1.39.7119

**Foote, J.** (2000). Automatic audio segmentation using a measure of audio novelty. In *Proc. of the IEEE International Conference on Multimedia and Expo*, volume 1, pages 452–455. DOI: https://doi.org/10.1109/ICME.2000.869637

**Foote, J. T.,** & **Cooper, M. L.** (2003). Media segmentation using self-similarity decomposition. In *Proc. SPIE 5021, Storage and Retrieval for Media Databases 2003*, pages 167–176. DOI: https://doi.org/10.1117/12.476302

**Grosche, P., Müller, M.,** & **Kurth, F.** (2010). Cyclic tempogram: A mid-level tempo representation for music signals. In *Proc. of the IEEE International Conference on Acoustics Speech and Signal Processing*, pages 5522–5525. DOI: https://doi.org/10.1109/ICASSP.2010.5495219

**Gulati, S.,** & **Rao, P.** (2010). Rhythm pattern representations for tempo detection in music. In *Proc. of the First International Conference on Intelligent Interactive Technologies and Multimedia*, pages 241–244. DOI: https://doi.org/10.1145/1963564.1963606

**Hermes, D. J.** (1990). Vowel-onset detection. *Journal of the Acoustical Society of America*, *87*(2), 866–873. DOI: https://doi.org/10.1121/1.398896

**Jensen, K.** (2006). Multiple scale music segmentation using rhythm, timbre, and harmony. *EURASIP Journal on Advances in Signal Processing*, *2007*(1), 1–11. DOI: https://doi.org/10.1155/2007/73205

**Jensen, K., Xu, J.,** & **Zachariasen, M.** (2005). Rhythm-based segmentation of popular Chinese music. In *Proc. of the International Conference on Music Information Retrieval*, pages 374–380.

**Klapuri, A., Virtanen, T., Eronen, A.,** & **Seppänen, J.** (2001). Automatic transcription of musical recordings. In *Proc. of the Consistent & Reliable Acoustic Cues*

*Workshop*. Aalborg, Denmark. DOI: https://doi.org/10.1109/ICCIMA.2007.138

**Kumar, P. P., Rao, P.,** & **Roy, S. D.** (2007). Note onset detection in natural humming. In *Proc. of the IEEE International Conference on Computational Intelligence and Multimedia Applications*, volume 4, pages 176–180.

**Logan, B.** (2000). Mel frequency cepstral coefficients for music modeling. In *Proc. of the International Symposium on Music Information Retrieval*.

**McCallum, M. C.** (2019). Unsupervised learning of deep features for music segmentation. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 346–350. DOI: https://doi.org/10.1109/ICASSP.2019.8683407

**McFee, B., Nieto, O., Farbood, M. M.,** & **Bello, J. P.** (2017). Evaluating hierarchical structure in music annotations. *Frontiers in Psychology*, *8*, 1337. DOI: https://doi.org/10.3389/fpsyg.2017.01337

**Paulus, J., Müller, M.,** & **Klapuri, A.** (2010). State of the art report: Audio-based music structure analysis. In *Proc. of the International Society for Music Information Retrieval Conference*, pages 625–636.

**Peeters, G.** (2003). Deriving musical structures from signal analysis for music audio summary generation: "Sequence" and "state" approach. In *Proc. of the International Symposium on Computer Music Modeling and Retrieval*, pages 143–166. DOI: https://doi.org/10.1007/978-3-540-39900-1_14

**Peeters, G.** (2007). Template-based estimation of timevarying tempo. *EURASIP Journal on Applied Signal Processing*, *2007*(1), 158–171. DOI: https://doi.org/10.1155/2007/67215

**Peeters, G.,** & **Deruty, E.** (2009). Is music structure annotation multi-dimensional? A proposal for robust local music annotation. In *Proc. of 3rd Workshop on Learning the Semantics of Audio Signals*, pages 75–90.

**Ranganathan, S.** (2013). Compositional models and aesthetic experience in dhruvapada. *The Music Academy Journal*, 84.

**Ranjani, H.,** & **Sreenivas, T.** (2013). Hierarchical classification of Carnatic music forms. In *Proc. of the International Society for Music Information Retrieval Conference*.

**Serra, X.** (2011). A multicultural approach in music information research. In *Proc. of the 12th International Society for Music Information Retrieval Conference*, pages 151–156.

**Smith, J. B. L., Burgoyne, J. A., Fujinaga, I., De Roure, D.,** & **Downie, J. S.** (2011). Design and creation of a large-scale database of structural annotations. In *Proc. of the International Society for Music Information Retrieval Conference*, pages 555––560.

**Srinivasamurthy, A., Holzapfel, A.,** & **Serra, X.** (2014). In search of automatic rhythm analysis methods for Turkish and Indian art music. *Journal of New Music Research*, *43*(1), 94–114. DOI: https://doi.org/10.1080/09298215.2013.879902

**Sundberg, J.** (1990). What's so special about singers? *Journal of Voice*, *4*(2), 107–119. DOI: https://doi.org/10.1016/S0892-1997(05)80135-3

**Thoshkahna, B., Müller, M., Kulkarni, V.,** & **Jiang, N.** (2015). Novel audio features for capturing tempo salience in music recordings. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 181–185. DOI: https://doi.org/10.1109/ICASSP.2015.7177956

**Tian, M.,** & **Sandler, M. B.** (2016). Towards music structural segmentation across genres: Features, structural hypotheses, and annotation principles. *ACM Transactions on Intelligent Systems and Technology*, *8*(2), 23. DOI: https://doi.org/10.1145/2950066

**Turnbull, D., Lanckriet, G. R., Pampalk, E.,** & **Goto, M.** (2007). A supervised approach for detecting boundaries in music using difference features and boosting. In *Proc. of the International Conference on Music Information Retrieval*, pages 51–54.

**Ullrich, K., Schlüter, J.,** & **Grill, T.** (2014). Boundary detection in music structure analysis using convolutional neural networks. In *Proc. of the International Society for Music Information Retrieval Conference*, pages 417–422.

**Verma, P., Vinutha, T. P., Pandit, P.,** & **Rao, P.** (2015). Structural segmentation of Hindustani concert audio with posterior features. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 136–140. DOI: https://doi.org/10.1109/ICASSP.2015.7177947

**Vinutha, T. P., Sankagiri, S., Ganguli, K. K.,** & **Rao, P.** (2016). Structural segmentation and visualization of sitar and sarod concert audio. In *Proc. of the International Society for Music Information Retrieval Conference*, pages 232–238.

**Wade, B. C.** (2001). *Music in India: The classical traditions*, Chapter 7: Performance Genres of Hindustani Music. Manohar Publishers.

**Widdess, R.** (1994). Involving the performers in transcription and analysis: A collaborative approach to dhrupad. *Ethnomusicology*, *38*(1), 59–79. DOI: https://doi.org/10.2307/852268

**Widdess, R.** (2011). Dynamics of melodic discourse in Indian music: Budhaditya Mukherjee's ālāp in rāg pūriyā-kalyān. In M. Tenzer & J. Roeder (Eds.), *Analytical and Cross-Cultural Studies in World Music*, pages 187–224. Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780195384581.003.0005

**Widdess, R.** (2013). Schemas and improvisation in Indian music. In R. Kempson, C. Howes & M. Orwin (Eds.), *Language, Music and Interaction*, pages 197–209. College Publications.

**How to cite this article:** Rao, P., Vinutha, T. P., & Rohit, M. A. (2020). Structural Segmentation of Alap in Dhrupad Vocal Concerts. *Transactions of the International Society for Music Information Retrieval,* 3(1), pp. 137–152. DOI: https://doi.org/10.5334/tismir.64

**Submitted:** 05 May 2020 **Accepted:** 07 July 2020 **Published:** 16 September 2020



]u[ *Transactions of the International Society for Music Information Retrieval* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS